



Methods of collecting facebook data and their effects on later analysis

Sandvik, Kjetil; Brügger, Niels; Laursen, Ditte

Publication date:
2013

Document version
Early version, also known as pre-print

Citation for published version (APA):
Sandvik, K., Brügger, N., & Laursen, D. (2013). *Methods of collecting facebook data and their effects on later analysis*. Paper presented at NORDMEDIA, Oslo, Norway.

Methods of collecting facebook material and their effects on later analyses

Ditte Laursen, Senior researcher
State Media Archive, State and University Library

Niels Brügger, Associate professor
Department of Aesthetics and Communication, Aarhus University

Kjetil Sandvik, Associate professor
Department of Media, Cognition and Communication, University of Copenhagen

This paper is work in progress; please do not quote without the authors' permission

Abstract

In this paper, we discuss how methods of archiving data on social network sites such as facebook necessarily reflect deletions (or selections) brought on by technical arrangements and theoretical interests. While collecting facebook data can be accomplished in a number of ways, eg. user diaries, observational notes etc., we discuss in particular the process of archiving the data from the web for analysis and documentation. Web data created in and by the process of archiving can be considered 'reborn' digital material, ie. an unique version of the material which we can never expect to find in the form it actually took on the web (Brügger, 2010a, 2011). Web data created in and by the process of archiving can also be considered both "found", ie. created through unobstrutive measures, and "made" insofar as they require a series of methodological choices and considerations to be meaningfully implemented in a research process (Jensen, 2012; Lomborg, 2012, cf. Brügger 2005). Furthering this body of research, we show in what ways four specific methods of web archiving each in its own way provide both limited and privileged access to the original data and, accordingly, frame what research questions can be answered. The four methods are 1) still image, 2) screen filming, 3) harvesting via API, and 4) web crawling. In addition, it is discussed how broad webarchiving in national web archives affect the possibilities for writing historical studies of, for instance, Facebook in the future, just as the interplay between internet and television archives is debated. In order to illustrate and feed the discussion, we build on an ongoing case study in which we have used all three methods. The case study is about the interplay between real-time internet and live television in the Danish talent programme Voice (Laursen, Sandvik, & Brügger, 2012).

Introduction

Social media has become increasingly prevalent in people's lives and also an important source for scholars to understand our time. Especially facebook, with more than 1 billion monthly users (Facebook, 2013), has attracted researchers' attention, and in recent years hundreds of academic articles have been published on facebook. Several literature reviews of the research about facebook have been published, mapping out themes and topics in the research (Di Capua, 2012; Graham, Gosling, & Wilson, 2012; Pérez-Latre, Portilla, & Sánchez-Blanco, 2012).

As any other social and cultural mediated phenomenon facebook can be studied in a number of ways using a variety of methods, be that methods from the humanities or the social sciences bringing into focus, for instance, image use,

discourse, argumentation, social interaction, friend networks, privacy, as well as the history of facebook, just to mention a few. However, despite the great amount of facebook studies, the methodologies of the studies have attracted only little attention.

According to one literature review of the research about facebook published 2007-2012 (Di Capua, 2012), the methodologies employed in most of the studies consist of surveys. Thus, in these studies, researchers preferred a quantitative approach and based their studies on self-report. According to a comprehensive literature search on 412 academic articles on facebook (Graham et al., 2012), there are three principle methods to collect facebook data: recruitment of participants in offline contexts, recruitment of participants via facebook applications, and data crawling. One review of studies using Facebook, Twitter and Youtube as datasources weigh strength and weaknesses of ethographical approaches, statistical approaches and computational approaches, calling for a mixed method approach (Giglietto, Rossi, & Bennato, 2012).

Apart from these studies, methodological reflections are typically bound to the specific method of the work in which it is deployed. For instance, Gjoka et al. consider and compare several techniques to crawl the social graph of facebook, weighing pros and cons of each (Gjoka, Kurant, Butts, & Markopoulou, 2011). Or, online methods are compared and weighed against traditional off-line methods (Bechmann & Vahlstrup, 2013; Rotman, Preece, He, & Druin, 2012).

In this paper, we wish to extend our knowledge with regard to methodological challenges in studying facebook. In particular, we will consider the methodological challenges which any scholar will encounter when facebook in its online form is transformed to and preserved as an object of study, in other words: When facebook is archived. Specifically, we will show in what ways four different methods of archiving each in its own way provide both limited and privileged access to facebook in its original form and, accordingly, frame what research questions can be answered. The four methods are 1) still image, 2) screen filming, 3) harvesting via API, and 4) web crawling. These methods are not meant to be an exhaustive list of methods or generalize toward all cases that comprise these methods. Instead, they will exemplify how different archiving techniques lend themselves to specific types of research questions. In order to illustrate and feed the discussion, we build on an case study in which we have used all four methods (Laursen et al., 2012). This case allows us to get at a process, which is context specific, and will demonstrate what original facebook material collected in four different ways looks like.

The paper begins by laying out different approaches to archiving and known methodological challenges in web archiving, including how the archived record is shaped by its context of creation. It will then go on in four sections which each correspond with the four archiving methods: still image, screen filming, harvesting via API, and web crawling. In each section, we will discuss the pro and cons of the specific method, drawing on our case study. In addition, we will refer to studies, which have used the method or discussed how the method's archived material can be analyzed.

Why archive web material?

- Need for a stable object. The transformational character of the web: The importance of the web makes it necessary to record and save copies to examine (Mitra & Cohen, 1999)
- documentation
- illustration

Main characteristics of archived web material

Historically, archives and scholars have dealt primarily with objects that are intuitively graspable from observation. Digital information, on the other hand, is more difficult to intuitively grasp, and in many cases impossible to grasp. The digital turn has forever changed the ways and means by which archives and scholars collect and archive objects (Calhoun, 2013; Dobрева & Ivacs, 2014; Foster & Rafferty, 2014; Marshall & Fieldhouse, 2012; Moss, 2008; Stevenson, 2008).

Found vs. made archived web material

- Jensen distinguishes between “found” and “made” data to delineate differences between data that are produced and archived in digital media regardless of a researcher’s interest and data that are produced and archived in the interaction between a researcher and a topic of interest (Jensen, 2011, p. 49). Most facebook studies are based on “made” web material. To our knowledge, only one study uses “found” facebook web material (Brügger, 2013).

Broad web archiving (macro archiving (Brügger, 2005))

- archiving institutions, such as national libraries
- aiming at preserving the cultural heritage of, for instance, a nation state
- allows for as many different kinds of research projects as possible in the future

The broad web archive is multitemporal and multispatial (Brügger & Finnemann, 2013)

- Since a web archive usually covers more than one point in past time, numerous versions of the same web element will exist — a URL, a web page, an image, a website, a hyperlink etc. — each from a different point in time
- Most often websites (especially larger websites) are not continuously archived in their totality — the spatial extension of the same archived website is not necessarily identical throughout time

Scholars working with web archives that have been collected as part of a general library or archive will need to address issues of sampling and boundaries, representation and selectivity, and archival standards (Schneider & Foot, 2008)

Narrow web archiving (micro archiving (Brügger, 2005))

- individual scholars or groups, in relation to, for instance, a specific research project
- usually calibrated to fit the research project in question

Digital archiving is "lifting"

- Reuse of a digital object requires "lifting" it out of its original context, then making sense and use of it in a new context (Guha & McCarthy, 2003)
- Archiving tools are not neutral, but leave material evidence of socio-cultural discourses (Fuller, 2008; Gillespie, 2003; Mackenzie, 2006)
- Unlike paper, which can be saved in their original form of presentation, the process of archiving is reflected in the artifact (Brügger, 2008)

The archived web is a reborn, unique and deficient version and not simply a copy of what was once online (Brügger, 2009; Brügger & Finnemann, 2013)

- The archiving institution or the researcher must make a number of choices
- What is archived is almost never a copy on a 1:1 scale of what was once online
 - it is not unusual for the website to change during the process of archiving, so something may be missing or there may be web elements in the archive which were not online at the same time
 - the archiving process may not have been performed as planned
 - once archived the website must be assembled and can be considered a reconstruction. It is created in and by the process of archiving, which is why it can be considered 'reborn' digital material, in contrast to digitized and born-digital material (cf. Brügger, 2011)

Analytical and methodological consequences

Inherent limits of lifting and the problem of context (Lee, 2007, 2011)

- No digital object can carry all of its context along with itself, ie. the object is never fully self-contained, self-describing
- Representation information can reside within the digital object itself, stored separately as metadata, and encoded within software required to read and parse the digital object
- Without access to sufficient contextual information, a user can suffer from gaps in understanding, but also, based on the propensity to make sense and reduce cognitive dissonance, incorrectly "fill in the gaps" based on characteristics of current context

The peculiarities of the archived web material makes it very difficult to determine what is missing (Brügger, 2009, 2012)

- No stable original to compare with
- Incompleteness is rarely documented

“in many cases a web archive is incomplete in such a way that it is hard to determine if something is missing at all, and if so, what and where. Since these shortcomings are an inherent part of the process of archiving, the archived website mostly does not communicate or document these, and we usually do not have other sources to indicate what might be missing” (Brügger, 2012)

Observation of archived digital data is indirect

- One cannot directly observe digital data, instead one can only observe the data displayed on a monitor or other output device (Carrier, 2006)
- Since the subdivision of the archived material and the subsequent combination of elements are not necessarily inscribed in the material itself, any 'montage' of the archived elements is also an editing of these elements, made a posteriori by either the web archive or the scholar (Brügger & Finnemann, 2013)

Data

The archived facebook material for this paper was gathered in relation to a research project about the synchronous interplay between television and the Internet in the final of the talent programme *Voice* on the Danish television station TV2 Danmark, November 24th, 2012 (Laursen et al., 2012; Sandvik & Laursen, 2013). The final show was produced on television and on the Internet simultaneously, and during the television show viewers were invited to engage with a facebook page. On this facebook page viewers could read posts and comment on the show. Moreover, they could follow the television production backstage in a live stream, and they could video chat with participants in the show. Productions like *Voice* are related to audience's growing tendency to use a second screen while watching television (nielsenwire, 2011, 2012).

For the purpose of this research project, we gathered the same facebook material in different ways. Initially, we did this as a solution to the different archiving problems we knew existed with these kinds of web material. While the national webarchive in Denmark do harvest facebook pages with Danish content, we knew that they would not be able to collect the streamed material. Consequently, we used filming software to supplement the data from the national webarchive. We also knew that the national archive would have problems in displaying all the comments on the facebook page. However, one of us was involved in a development project on data extraction via facebook's API, and we were able to extract all the data from the facebook page with software from this project. In this way, we archived the same webmaterial in three ways: via filming, data extraction from API, and web harvesting. We supplemented this material with still images of the facebook page.

[overgang]

Still image

A screen shot or a screen dump is a record in the form of an image, displaying the visible items on the computer screen. Usually, the image is taken by the computer operating system or software running on the computer. Most applications can capture a chosen area, a whole window or the whole screen. On [Mac OS X](#), a user can take a screenshot of an entire screen by pressing ⌘ Cmd + ⇧ Shift + 3, or of a chosen area of the screen by ⌘ Cmd + ⇧ Shift + 4. The tiff-file will automatically be saved in a specified folder.

The screen shot provides a look a like version of the facebook page, see figure 1 and figure 2. In other words, the researcher will get a good overall sense of what the webpage looked like on the web. The tiff-format is also a robust format, which can be read by a range of softwares.

However, it is a cropped version, capturing only what is displayed on the computer screen. The researcher is not able to scroll up or down or to the sides to see the full page. Moreover, the text on the page is not clickable. This means that the hyperlink does not function as a hyperlink, eg. as a mean to move to other parts of the displayed page or to other pages linked to the displayed page. In addition, the screenshot is not machine-readable. As a consequence, the screenshot is not searchable. Finally, the streamed video content has lost its movements and sounds when it was turned into one single image.

For researchers interested in the visual appearance of the page, the screen shot may be a suitable solution. For instance, the page layout can be identified and mapped out, including the potential culturally specific meanings that reside in the visual properties (cf. Pauwels, 2011, 2012). It is possible to perform a syntactical analysis of interrelated textual elements, mapping out weak or strong relations between elements and hierarchies between elements (cf. Brügger, 2007; Brügger, 2010b). And it is possible to analyze the page's style (cf. Engholm, 2002) and how the page communicates through its graphical design, content design, interaction design and social design (cf. Engholm & Klastrup, 2010). In short, it is possible to study *how* the page communicates.

For researchers interested in *what* the page communicates, the screen shot may not be the best choice. While one can get an overall sense of the content and the activity level on the facebook wall, an analysis of the content (Herring, 2010), qualitative or quantitative (Scheufele, 2008a, 2008b), would be more manageable in a text format. In the image of the facebook live chat, the content is completely lost (figure 2).

In our review of facebook studies, we have only come across one researcher, who states he has used screenshots as primary analytical objects. In his history of facebook's website 2004-2013, Brügger used screenshots found on the open web to analyze the website's features and semiotic and interactional elements over time (Brügger, 2013). A more common use of screenshots, however, is to document and illustrate a study. For instance, a qualitative analysis of a facebook page as rhetorical discourse is documented by given the reader a screenshot of the page analyzed (Agerbæk, 2008). Similarly, shotshots are provided in relation to websites analyzed by Thorlacius, with her model for analyzing visual communication in web design (Thorlacius, 2002, 2010). In both cases, it is clear from the analyses that they are not

solely based on the screenshots. In several studies, the screenshot serves as mere illustration. For instance, a study of wall posts during the Thailand floods in 2011 uses a screenshot to display the facebook page from where the posts were collected (Kim et al., 2013; see also Krutern, 2012). In another example, a study based on two quantitative surveys on the construction of digital identities on facebook is illustrated with screenshots from the author's own facebook account (Strano, 2008).

advantages:

- look a like version
- robust format

disadvantages:

- cropped version
- not machine-readable
- not clickable
- no moving images
- no sound



Figure 1: Screenshot of facebook page

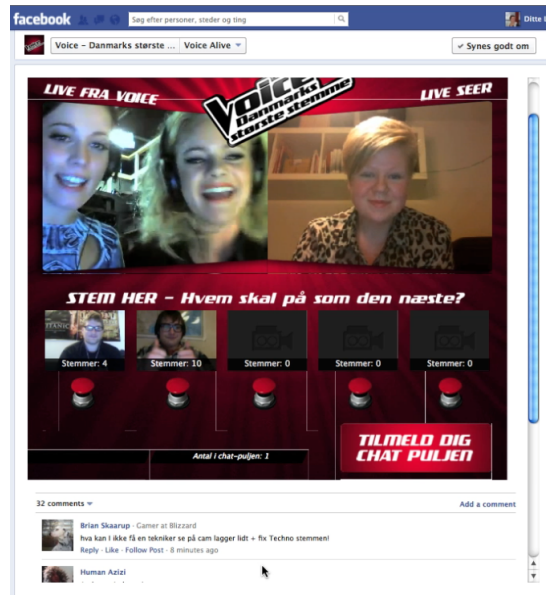


Figure 2: Screenshot of facebook live videochat page

An alternative to making a screenshot is to convert the facebook page into a pdf-file. Add-ons like Fireshot or services like pdfmyurl.com allow for capture of entire webpages. In some applications, it is also possible to capture a selection or just a visible part of the webpage. The pdf is basically a conversion of the html page. All the applications differ in their output, and the researcher may have to try several applications to find one, which serves his needs the best.

The conversion creates a representation, which has several advantages. Unlike the screenshot, the pdf is a version of the entire page. Moreover, it is allowing for searches. However, pdf files are typically difficult for machines to interpret. In addition, as evident in figure 3 and 4, the representation is not nearly as close to the original webpage as the basic screenshot. And while the researcher will get a capture of the entire page, content is lost in process. For instance, the facebook wall seems cropped, and sounds and moving images are not captured. Finally, not all generated pdf's are clickable, eg. the multiple levels of the page is not detained, only the level of the specific URL.

For researchers interested in some kinds of content analysis, the pdf may be a suitable solution. While some of the content is lost in the conversion, the pdf of the facebook wall page will provide text versions of, for instance, all status updates. However, since the original links in the pdf version typically are not clickable, the researcher is not able, for instance, to see all the comments. In the same way, the researcher can study the photos and their context, but not, for instance, the video stream.

In our review of facebook studies, we have not come across researchers, who state they have used pdf's as primary analytical objects. Nor have we come across researchers who have documented or illustrated their study with pdf's.

advantages:

- capture of entire page
- searchable
- robust format

disadvantages:

- not clickable
- not machine-readable
- no moving images
- no sound
- distorted version



Voice – Danmarks største stemme

Synes g

23.474 Synes godt om · tilkendegivelser · 4.914 taler om dette

Tv-program

Den officielle Facebook side for 'Voice – Danmarks største stemme'. Følg med på www.v.tv2.dk

Billeder

Voice Alive

Synes godt om

Om

Højdepunkter

Opslag

Billede/video

Skriv noget...

Voice – Danmarks største stemme delte et link, for omkring en time siden

Selvom Andreas måtte se sig slået af Emilie i Voice-finalen, så er hans selvtillid intakt, og han vil lave den bedste danske plade... nogensinde.

Voice – Danmarks største stemme – "Jeg laver den bedste danske plade nogensinde" v.tv2.dk

2 Venner

synes godt om Voice – Danmark



Seneste opslag af andre

Zanne Hansen

Hvor er den top fra som Sha 1 · for omkring en time s

Bente Berg Andersen

1 of 14

25/11/12 21.0521

Figure 3: First page of a generated pdf of the facebook wall

10

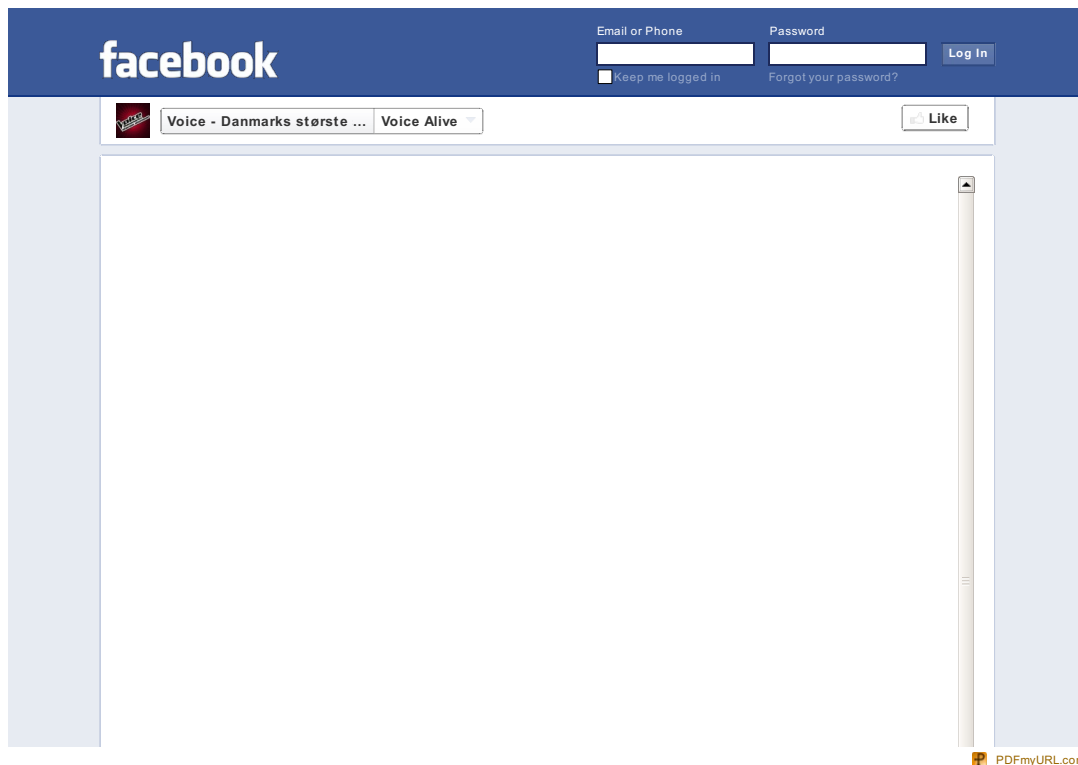


Figure 4: A generated pdf of the facebook live video chat

Screen filming

The screen filming is a video capture of the computer screen output, and it can include audio. Applications like SnapzPro and Camstudio allow for video capture. Most applications permit the researcher to record a particular area on the screen or a program window so that the rest of the desktop is not showing on the recording. Most applications can also record at different frame rate speeds, for example 1 FPS (frames per second) to create a timelapse video effect, or 30 FPS for a smooth video. Videos are typically produced in AVI or MOV file format. Some programs have options to enable or disable the mouse cursor, record sounds from programs or a microphone (or have no sound at all) and the option to enable custom screen annotations.

As with the still image, the screen filming provides a look a like version of the facebook page, see figure 5 and 6. In addition, screen filming captures a version of an event as it happens. In other words, it allow researchers to get closer to the experiences of the recipients in providing oppurtunities to record various aspects of social practices in real time: talk, bodily conduct, material environments etc. The amount of detail that can be captured in video recordings makes them a powerful resource. In addition, unlike other forms of social scientific data, there is an opportunity to pause, rewind and fastforward, thus reframing the analytic gaze.

Nevertheless, the screen filming has several disadvantages in common with the still image: It is a cropped version, and it is not searchable or clickable. Moreover, it is not machine-readable, at least not for the common researcher.

For researchers interested in movements and sounds, screen filming could be a good choice. Thus, the livestreamed video chat from our case is entirely captured in the movie (figure 6), including the audio. In consequence, the recording enables all sorts of video and audio analysis of interactions and events. However, not all movements and sounds can be captured automatically by screen filming software. A video posted on the facebook wall does not play by itself. In such a case, the researcher or a participant in the study would have to use the mouse to play it while the software is running.

Also for researchers interested in the moment-by-moment development of a facebook wall and the timing of events, screen filming could be a good choice. Since facebook pages, unlike other kind of websites, are not stable for sometimes longer than a few seconds or minutes, capturing them in their temporal evolution becomes all the more central. With our three-hour long video of the facebook wall during the finals, we were able to follow development over time, including tendencies in postings and comments and activity levels. With snapshots by screen capturing software, the capture of the moment-by-moment development would not have been feasible. Additionally, snapshots would evidently have captured events at a specific moment in time, not all moments.

However, in order to capture moment-by-moment development on the wall without an interacting user, the researcher would have to enable an automated reload of the page. In our case, we downloaded the 'reload every' add-on and set it up to reload the page automatically every 5 minutes. In doing so, we did in fact create a series, which captured most of the activities, but not all. When the winner of the television show was announced on the facebook wall at 10:40 pm, the activity level was so high that not every comment was captured. This is evident in figure 6, where the winner is announced 'about a minute ago', and 35 comments have appeared while only one has been displayed in full. This reminds us, again, that video recordings are a version of the event and cannot be ultimately accorded objective status. Like other forms of archived data, video recordings selectively delete or foreground aspects of the original data. The recordings are brought on by technical arrangements and theoretical interests afforded by the software and expressed in the researcher's choices in relation to the use of the software (frame, framerate, reload speed, mouse movements etc.).

The use of film as data in the social sciences has a long and distinguished history (Heath, Hindmarsh, & Luff, 2010; Knoblauch, Schnettler, & Raab, 2012). Video-based research has also been applied to webpages, especially virtual environments (Boellstorff, Nardi, Pearce, & Taylor, 2012; Consalvo & Dutton, 2006; Hine, 2005). However, in our review of the academic literature on facebook, we have found only one study, which uses screen filming. Meredith and Potter (2014) used screen filming of their facebook users' computers, and showed that, and how, instant messages are edited before they are sent and equally how new messages come in while initial messages are still being typed. Without the screen filming, the two researchers would not have been able to analyze the temporal and sequential

structure of the micro events on the facebook page, even less to document and illustrate their findings.

advantages:

- look a like version
- robust format
- moving images
- sound
- development over time
- timing of events

disadvantages:

- cropped version
- not machine-readable
- not clickable

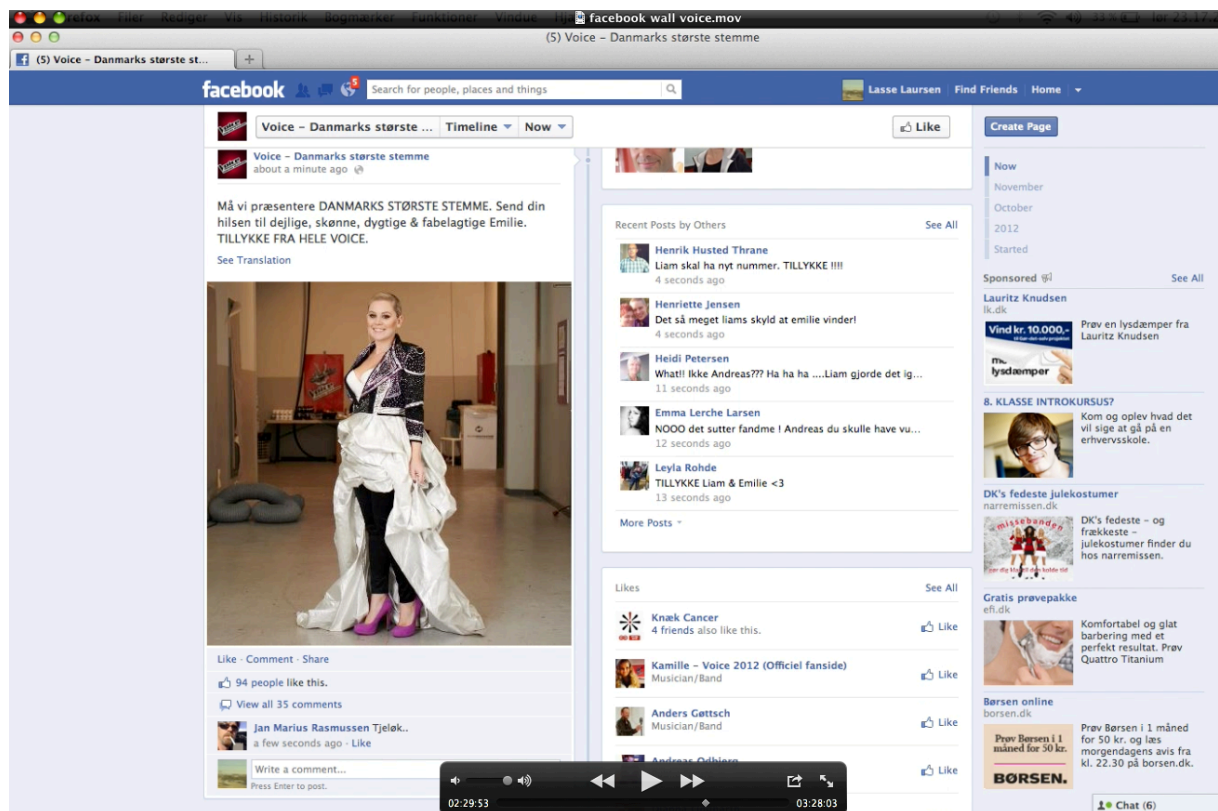


Figure 5: A screen movie of the facebook wall



Figure 6: A screen movie of the facebook live video chat

Harvesting via API

Harvesting a site via API means using the html-based open application-programming interface (API) of the site's server to collect its content. The API is a programmatic interface to a defined request-response message system, typically expressed in JSON or XML. [...] API harvesters are typically customized in different ways. We used a system called Digital Footprints (digitalfootprints.dk).

[...]

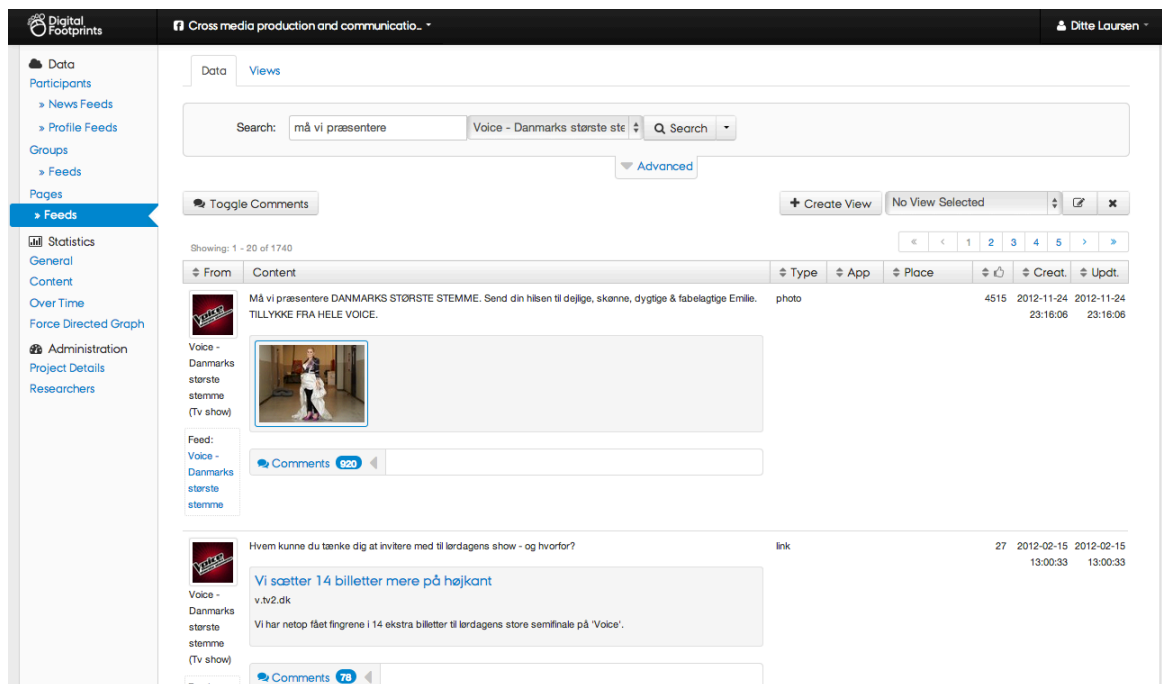


Figure 7: Front end of database with material harvested via API



Figure 8: Post by user, harvested via API

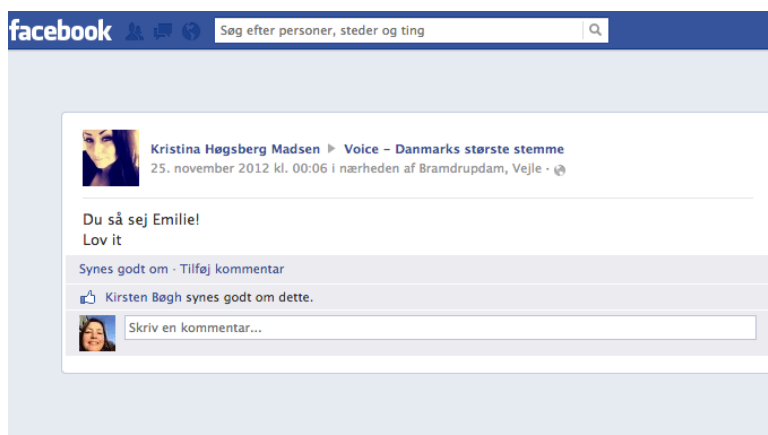


Figure 9: Still image of post (compare with figure 8)

advantages:

- capture of entire page's content
- capture of developments over time
- machine-readable (searchable, clickable, sortable)
- access to otherwise hidden data
- possible to manage data as big data

disadvantages (some are specific for this particular software):

- no capture of the original visual display of elements
- videos are not playable (instead, link to live web)
- no capture of streamed content
- no capture of linked-to content (instead, forward to live web)
- proprietary format

Web crawling

Web crawling is the process of collecting web material and loading it into an a fully browsable web archive, with working links, media etc. A web crawler starts with a domain name or list of URLs to visit. As the crawler crawls, it identifies new links and is able to crawl these sites as well. The crawlers also archive metadata about the collected resources such as access time and content-length. Using web crawlers requires technical skills and are usually done by large institutions such as national libraries. One of biggest native format web archive is the Internet Archive (archive.org), which strives to maintain an archive of the entire Web. In Denmark, the national webarchive (netarkivet.dk) has collected facebook material by web crawlers since 2005. There are numerous services that may be used to archive web resources on-demand, using web crawling techniques, ie. Archive-It (archive-it.org).

[...]

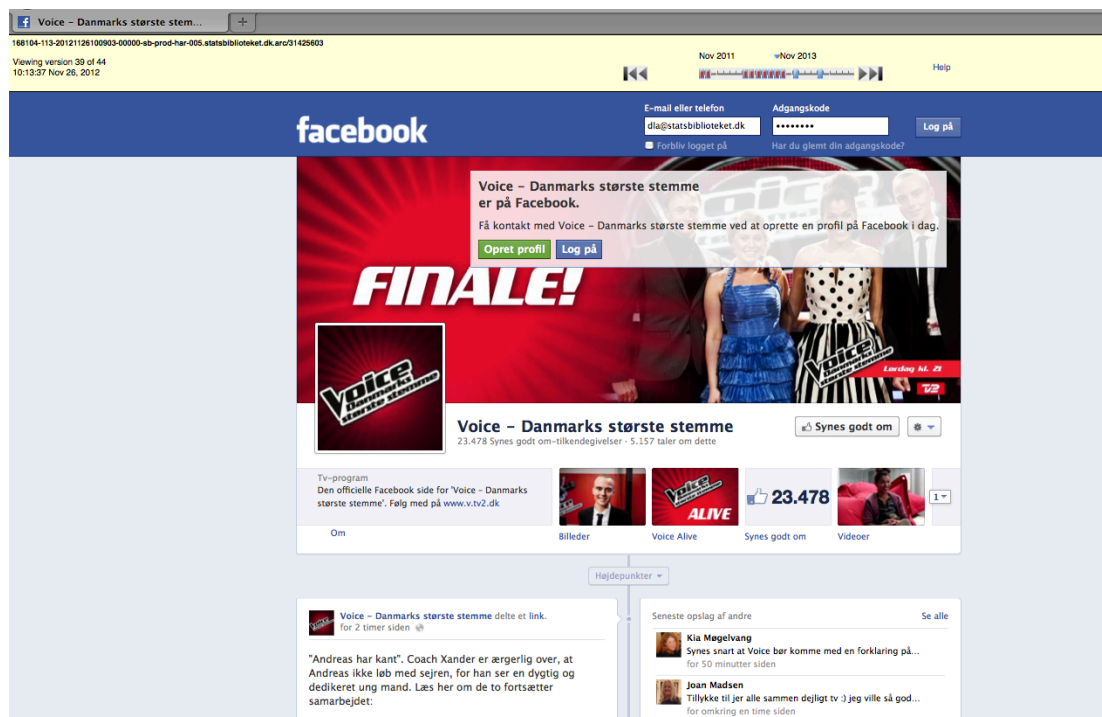


Figure 10: Facebook page captured by the national netarchive's web crawlers Nov 26, 2012, two days after the finals.

Voice – Danmarks største ...
Tidslinje
Nu

Voice – Danmarks største ...

lørdag kl. 15:48
Bøje Andreassen Skat dit dejlige hug fortalte jo at der er fusk i det her lort... tøsen får max udgivet en cd
lørdag kl. 17:08

Voice – Danmarks største stemme
lørdag

Må vi præsentere DANMARKS STØRSTE STEMME. Send din hilsen til dejlige, skønne, dygtige & fabelagtige Emilie. TILLYKKE FRA HELE VOICE.

Synes godt om · Tilføj kommentar · Del

64

Marianne Langkjær Sondergaard, Linette Grønborg, Julie Sjøgren og 4.487 andre synes godt om dette.

Vis tidligere kommentarer
2 af 942

Daniella Lenskjold

tillykke tillykke flot 😊

for 4 timer siden

Martin Thilm

congrats, very good, fyi 😊

Se oversættelse

for 2 timer siden

Voice – Danmarks største stemme
lørdag

Maria Holmen

Har længe holdt med Andreas, men syntes ærligt talt ikke han var på toppen i aften, hans første sangvalg passede ikke helt rigtigt og han fik ikke vist hvad det er der gør ham speciel, så baseret på i aften var Emilie en sand vinder! Men kun 2 gange i Voice har jeg tænkt WOW, begge optrædender var Andreas 😊

lørdag kl. 17:02 · 1

Pia Johannessen

Emilie, du var ikke min favorit, men du gjorde det godt, og har fortjent at vinde. Christian var min favorit. Ps. Er der nogen der stemmer efter hvem de ha som choastes!!!!

lørdag kl. 22:37 via mobil · 1

Voice – Danmarks største stemme
lørdag

Endnu en dygtig og dejlig sanger må forlade finalen. Vi har nydt dit selskab hver eneste gang. Giv ham en hånd.

Synes godt om · Tilføj kommentar · Del

12

Jeanie Schuster, Vibeke Sørensen, Mette Levisen og 1.342 andre synes godt om dette.

Vis tidligere kommentarer
2 af 278

Figure 11: By scrolling down, material on the finals can be found on Facebook page captured by the national netarchive's web crawlers Nov 26, 2012, two days after the finals.



Figure 12: Facebook live video chat captured by the national netarchive's web crawlers Nov 26, 2012

advantages:

- capture of entire page
- machine-readable (searchable, clickable, sortable)
- access to otherwise hidden data (logfiles)
- capture of linked-to content (if included in the crawling)
- possible to manage data as big data
- robust format (html)

disadvantages (some are specific for this particular software):

- look a like version (except for some applications, ie. flash, java script ...)
- temporal delimitation
- spatial delimitation (to some extent)
- no capture of streamed content
- no capture of videos

Overview

	Replica version (look a like)	One version	Entire version (not cropped)	Native formats version	Temporal version
Screen shot image	x	x			
Screen shot pdf		x	x		
Screen movie (wall)	x	x			x
Screen movie (live chat)	x	x			x
Data extract from API (wall)		x	x	x	
Web crawler (wall)	x		x	x	
Web crawler (live chat)	x		x	(x)	

Discussion

[...]

Conclusion

In this paper we have considered four different approaches to facebook archiving and the research opportunities they each offer to researchers. [...]

References

- Agerbæk, L. (2008). *Is a profile in social software a learning e-portfolio? If not, could any benefits be found from linking the two?* (Vol. 25).
- Bechmann, A., & Vahlstrup, P. B. (2013). Digital Footprints: A system for studying private user data on Facebook with consent. *CHI2013*.
- Boellstorff, T., Nardi, B., Pearce, C., & Taylor, T. L. (2012). *Ethnography and virtual worlds : a handbook of method*. Princeton: Princeton University Press.
- Brügger, N. (2005). *Archiving websites. General considerations and strategies*. Aarhus: The centre for Internet research.
- Brügger, N. (2007). The website as unit of analysis? Bolter and Manovich revisited. *Northern Lights*, 5(1), 75-88. doi: 10.1386/nl.5.1.75_1
- Brügger, N. (2008). The Archived Website and Website Philology: A New Type of Historical Document? *Nordicom Review*, 29(2), 155-175.
- Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1-2), 115-132. doi: 10.1177/1461444808099574
- Brügger, N. (2010a). Web history, an emerging field of study. In N. Brügger (Ed.), *Web History* (pp. 22-59). New York: Peter Lang.
- Brügger, N. (2010b). *Website Analysis. Elements of a conceptual architecture* (Vol. 12). Aarhus: The Centre for Internet Research.

- Brügger, N. (2011). Web archiving – between past, present, and future. In M. Consalvo & C. Ess (Eds.), *The handbook of internet studies* (pp. 22-42). Oxford, England: Wiley-Blackwell.
- Brügger, N. (2012). Web History and the Web as a Historical Source. *Zeithistorische Forschungen/Studies in Contemporary History*, 9(2).
- Brügger, N. (2013). Facebooks historie: Udviklingen af en tom struktur. In J. L. Jensen & J. Tække (Eds.), *Facebook i den danske hverdag — fra socialt netværk til metamedie*. København: Samfundslitteratur.
- Brügger, N., & Finnemann, N. O. (2013). The Web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media*, 57(1), 66-80.
- Calhoun, K. (2013). *Digital Libraries*. GB: Facet Publishing.
- Carrier, B. D. (2006). *A hypothesis-based approach to digital forensic investigations*: ProQuest.
- Consalvo, M., & Dutton, N. (2006). Game analysis: Developing a methodological toolkit for the qualitative study of games. *Game Studies*, 6(1), 1-17.
- Di Capua, I. (2012). A Literature Review of Research on Facebook Use. *Open Communication Journal*, 6, 37-42.
- Dobrev, M., & Ivacs, G. (Eds.). (2014). *Digital Archives: Management, use and access* GB: Facet Publishing.
- Engholm, I. (2002). Digital style history: the development of graphic design on the Internet. *Digital Creativity*, 13(4), 193-211. doi: 10.1076/digc.13.4.193.8672
- Engholm, I., & Klastrup, L. (2010). *Websites as Artefacts: a New Model for Website Analysis*. Paper presented at the Konferenssissa 2nd International Conference on New Media and Interactivity (NMIC) Istanbul, Turkki.
- Facebook. (2013). Facebook Reports First Quarter 2013 Results Retrieved July 10, 2013, from <http://investor.fb.com/releasedetail.cfm?ReleaseID=761090>
- Foster, A., & Rafferty, P. (Eds.). (2014). *Managing Digital Cultural Objects: Analysis, discovery and retrieval* GB: Facet Publishing.
- Fuller, M. (2008). *Software studies: A lexicon*. Cambridge: The MIT Press.
- Giglietto, F., Rossi, L., & Bennato, D. (2012). The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source. *Journal of Technology in Human Services*, 30(3-4), 145-159. doi: 10.1080/15228835.2012.743797
- Gillespie, T. (2003). The stories digital tools tell. In J. Caldwell & A. Everett (Eds.), *New Media: Theories and Practices of Digitextuality* (pp. 107-126). New York: Routledge.
- Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2011). Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications*, 29(9), 1872-1881. doi: 10.1109/jsac.2011.111011
- Graham, L. T., Gosling, S. D., & Wilson, R. E. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*, 7(3), 203-220. doi: 10.1177/1745691612442904
- Guha, R., & McCarthy, J. (2003). Varieties of Contexts. In P. Blackburn, C. Ghidini, R. Turner & F. Giunchiglia (Eds.), *Modeling and Using Context* (Vol. 2680, pp. 164-177): Springer Berlin Heidelberg.

- Heath, C., Hindmarsh, J., & Luff, P. (2010). Video, analysis and the social sciences. In C. Heath, J. Hindmarsh & P. Luff (Eds.), *Video in qualitative research : analysing social interaction in everyday life* (pp. 1-13). Los Angeles: SAGE.
- Herring, S. C. (2010). Web content analysis: Expanding the paradigm *International handbook of Internet research* (pp. 233-249). New York: Springer.
- Hine, C. (2005). *Virtual methods : issues in social research on the Internet*. Oxford: Berg.
- Jensen, K. B. (2012). Lost, found, and made: Qualitative data in the study of threestep flows of communication. In I. Volkmer (Ed.), *Handbook of global media research* (pp. 435-450). Malden, MA: Wiley-Blackwell.
- Kim, C., Kang, B. S., Choi, H. J., Lee, Y. J., Kang, G. H., Choi, W. J., & Kwon, I. H. (2013). Nationwide online social networking for cardiovascular care in Korea using Facebook. *Journal of the American Medical Informatics Association*. doi: 10.1136/amiajnl-2012-001465
- Knoblauch, H., Schnettler, B., & Raab, J. (2012). Video-analysis. Methodological aspects of interpretive audiovisual analysis in social research. In H. Knoblauch, -, H.-G. Soeffner, J. Raab & B. Schnettler (Eds.), *Video analysis: methodology and methods : qualitative audiovisual data analysis in sociology* (pp. 9-26). Frankfurt, M.: Lang.
- Kruter, V. (2012). Social network site, a citizen's mouthpiece in time of crisis-The study of Facebook use in Thailand floods 2011.
- Laursen, D., Sandvik, K., & Brügger, N. (2012). Cross media production and communication: Interplay between real-time internet and live television in the Danish talent programme Voice Retrieved February 18, 2013, from <http://www.netlab.dk/projects/p5-cross-media-production-and-communication/>
- Lee, C. A. (2007). Taking context seriously: a framework for contextual information in digital collections. *UNC SILS TR-2007-04*.
- Lee, C. A. (2011). A framework for contextual information in digital collections. *Journal of Documentation*, 67(1), 95-143.
- Lomborg, S. (2012). Researching Communicative Practice: Web Archiving in Qualitative Social Media Research. *Journal of Technology in Human Services*, 30(3-4), 219-231. doi: 10.1080/15228835.2012.744719
- Mackenzie, A. (2006). *Cutting code: Software and sociality* (Vol. 30). New York: Peter Lang.
- Marshall, A., & Fieldhouse, M. (2012). *Collection development in the digital age*. GB: Facet Publishing.
- Meredith, J., & Potter, J. (2014). Conversation Analysis and Electronic Interactions: Methodological, Analytic and Technical Considerations *Innovative Methods and Technologies for Electronic Discourse Analysis* (pp. 370-393): IGI Global.
- Mitra, A., & Cohen, E. (1999). Analyzing the Web. In S. Jones (Ed.), *Doing Internet research: Critical issues and methods for examining the Net* (pp. 179-202). Thousand Oaks, Calif.: Sage.
- Moss, M. (2008). Opening Pandora's Box: What is an Archive in the Digital Environment? In L. Craven (Ed.), *What Are Archives?: Cultural and Theoretical Perspectives: A Reader* (pp. 71). Aldershot, England, Burlington, VT: Ashgate.

- nielsenwire. (2011, May 19). In the U.S., Tablets are TV Buddies while eReaders Make Great Bedfellows Retrieved February 18, 2013, from http://blog.nielsen.com/nielsenwire/online_mobile/in-the-u-s-tablets-are-tv-buddies-while-ereaders-make-great-bedfellows/
- nielsenwire. (2012). Double Vision – Global Trends in Tablet and Smartphone Use while Watching TV Retrieved February 18, 2013, from http://blog.nielsen.com/nielsenwire/online_mobile/double-vision-global-trends-in-tablet-and-smartphone-use-while-watching-tv/
- Pauwels, L. (2011). Researching websites as social and cultural expressions: methodological predicaments and a multimodal model for analysis. In E. Margolis & L. Pauwels (Eds.), *Visual research methods* (pp. 570-589). Los Angeles: Sage.
- Pauwels, L. (2012). A Multimodal Framework for Analyzing Websites as Cultural Expressions. [Article]. *Journal of Computer-Mediated Communication*, 17(3), 247-265. doi: 10.1111/j.1083-6101.2012.01572.x
- Pérez-Latre, F. J., Portilla, I., & Sánchez-Blanco, C. (2012). Social networks, media and audiences: A literature review.
- Rotman, D., Preece, J., He, Y., & Druin, A. (2012). *Extreme ethnography: challenges for research in large scale online environments*. Paper presented at the Proceedings of the 2012 iConference.
- Sandvik, K., & Laursen, D. (2013). *Second Screen Production: Creating rich media experiences through synchronous interplay between TV, web and social media*. Paper presented at the Nordmedia Conference, Oslo.
- Scheufele, B. (2008a). Content analysis, qualitative. In W. Donsbach (Ed.), *The International Encyclopedia of Communication* Oxford: Blackwell.
- Scheufele, B. (2008b). Content analysis, quantitative. In W. Donsbach (Ed.), *The International Encyclopedia of Communication* (pp. 967-972). Oxford: Blackwell.
- Schneider, S. M., & Foot, K. (2008). Archiving of Internet Content. In W. Donsbach (Ed.), *The International Encyclopedia of Communication*. Oxford: Blackwell.
- Stevenson, J. (2008). The Online Archivist: A Positive Approach to the Digital Information Age. In L. Craven (Ed.), *What Are Archives?: Cultural and Theoretical Perspectives: A Reader* (pp. 89). Aldershot, England, Burlington, VT: Ashgate.
- Strano, M. M. (2008). User descriptions and interpretations of self-presentation through Facebook profile images. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 2(2), 5.
- Thorlacius, L. (2002). A model of visual, aesthetic communication focusing on web sites. *Digital Creativity*, 13(2), 85-98.
- Thorlacius, L. (2010). Visual communication in web design – analyzing visual communication in web design. In J. Hunsiger, L. Klastrop & M. Allen (Eds.), *The International Handbook of Internet Research* (pp. 455-476). New York: Springer.